Chapter 2

ESTIMATION OF THE REGULARITY OF RENTAL PAYMENTS BY CUSTOMERS IN THE REAL ESTATE INDUSTRY THROUGH DATA MINING

Müberra ŞEN[1]

INTRODUCTION

As companies' data ownership continues to increase, the importance of data is increasing day by day. Because innovations in technology develop solutions for the use of this complex and large amount of data. Companies work to get the most out of their data and drive results in their favour. Since the emergence of the Data Mining concept, many studies have been carried out in practice. This is crucial for the analysis and use of big data. Data and data mining has gained great importance in almost every sector today. These areas include real estate, healthcare, finance, retail, manufacturing, education and many more. Data mining consists of a set of statistical and mathematical techniques developed to process, analyze and extract meaningful information from large and complex data sets in these various industries. Technologies such as machine learning and artificial intelligence enable the effective use of data mining. In this way, it is possible to examine the data in depth and to discover hidden relationships and patterns. Therefore, in various industries, including real estate, data mining has become an indispensable tool for making better decisions, gaining competitive advantage, innovating and predicting future trends. In this study, the real estate sector was examined. Regular payment of rents is an important issue in the real estate field. Many homeowners struggle to rent out their homes and turn them into a source of income. Then encountering a tenant who does not pay their rent regularly creates great difficulties and problems for both landlords and real estate agents. This may lead to the termination of the lease agreement due to non-payment of the rental fee. Therefore, this has become a very common problem. An analysis

[1] Marmara University, Faculty of Management, German Business Informatics, Muberra.senn@gmail.com, ORCID iD: xxx

will be made with the information of the tenants held and an estimate will be made for the potential tenants. With the application of data mining related to this problem, it is aimed to bring the landlords together with the potential tenants who only pay regular rent. In other words, by designing a model and training this model with current tenant information, the estimation of the tenants who will pay the rent will be realized. In this case, the relationship between the demographic characteristics of the tenants in the regular payment of their rent will be examined. Is the tenant's regular rent payment dependent on gender, education, income and rental prices? Does pay status depend on demographics? Which factor is more important if there is a link with these features? Is it possible to predict whether a potential tenant will make regular payments based on information gathered from the current tenant? The purpose of this study is to make an assessment of whether the potential tenants of the house can pay the rent regularly. The goal is to ensure that new tenants pay their rent properly. These estimates provide important information for real estate investors and real estate companies. Regular rent payments provide the landlord with a stable source of income and make it easier to pay off their financial obligations. Therefore, it is important to evaluate the regularity of rental payments in order to minimize risks and be prepared for potential problems. Real estate companies, on the other hand, want to estimate the regularity of rental payments to predict the performance and revenues of their real estate portfolios. With the positive conclusion of the contracts they have made, it will expand its volume in the market. An estimate will be made for the potential tenant, taking into account some demographic characteristics of people who already have contracts with real estate companies as tenants.

The machine learning application to be created creates a model suitable for the desired purpose and aims to find the appropriate tenant by using the data at hand thanks to this model. When the application is made, companies in the real estate sector will also show great interest in this study and will gain many advantages compared to other companies by using this application.

DATA AND DATA MINING

This section first defines data mining and the meaning of data, emphasizing the goal of data mining to make complex data understandable. Data mining is explained as the process of extracting meaningful information from raw data using statistical, mathematical, artificial intelligence and database techniques. It discusses how this process is applied in the real estate industry with the impact of the big data concept. In addition, data mining models are divided into predictive and descriptive methods, with particular emphasis on predictive models and classification models. Then, the success factors and challenges of data mining are discussed, emphasizing that data mining requires the right model selection, data quality, diversity, and foresight. This section summarizes important issues in the field, including data mining and its applications, models and challenges in the real estate sector.

DEFINITION AND MEANING OF DATA AND DATA MINING

The term data mining is defined by Zülfikar (Zülfikar, 2022) as "the state of ready-to-use information before it is completed". It is designed to make it simple, clean and understandable by working with messy, irregular data. Because if data is not converted into information, it is useless and cannot be used. Because it is in a heap and no benefit can be obtained from it. Data can be obtained from any source imaginable (Zülfikar, 2022). What matters is the question of what data is needed when choosing a source. Without data and the information derived from it, progress is not possible in the age of technology.

Data mining is defined as a company's ability to obtain seamless and meaningful data from a large data inventory on the required subject. Data mining is a process that uses statistical, mathematical, artificial intelligence and database techniques to discover and extract meaningful information from raw data. This process usually takes place by analyzing large amounts of data. Here are the steps that explain the data mining process in more detail. This information should be supported at all operational stages of the company. By obtaining all the data needed and working on this basis, companies will be one step ahead of their competitors. Therefore, it would actually be a minus for companies to abandon today's data mining efforts. Many companies are aware of this situation and act accordingly (Huria, 2014).

Applications of Data Mining in Real Estate

In the real estate sector, the amount of data obtained with the concept of big data has increased. For this reason, studies using machine learning applications have been conducted in this area. Popular studies include real estate appraisal and price estimation taking into account certain criteria. Property appraisal studies have been conducted using machine learning tools to make estimates based on property characteristics. In studies, it is possible to estimate the current value of an existing home. Accordingly, it is advantageous for the homeowner to be able to sell at the right price. Looking at other studies, the results of the surveys enable citizens to find the desired apartment more easily and comfortably when choosing an apartment. In addition, an attempt was made to answer which factors influence people's life decisions. Research has been conducted into the rationale behind the selection of potentially habitable houses. Analyzes were performed to determine if the suspected factors had an impact. Such and similar studies have brought a new dimension to the real estate sector with data mining. If you look at the research, it is mostly used for prediction. Allows you to draw conclusions about marketing and reliability in the real estate sector (Akay et al., 2023).

Data Mining Models

Data mining models fall into two categories. Estimation method models are one of them. After that, they are divided into two areas: regression analysis and classification. Models for regression analysis include: Linear regression describes the functional state of the relationship between the independent variable and the dependent variable. Logistic regression is used when the target variables or independent dependent variables contain zeros and ones. This regression model can address problems that can be solved by categorization (Erden, 2020).

Classification models: It has a schematic structure of decision trees. The decision tree has the highest root. When divided downward, it forms branches and then leaves. These assignments are actually properties. The completion of this structure, which extends from the root to the leaves, leads to the desired result. In other words, when the leaves are reached, the factors that need to be predicted will also be realized (Çelik, 2009).

Bayesian Classification: Before starting the Bayesian technique, it is determined which variable has the highest probability in terms of probabilities. Probability evaluation of the learning data used to train the model is made. (Kodedu, 2014).

The nearest neighbor's K-classification defines the distance between the target data value and the k-value. Euclidean distance is often used to calculate the distance to neighbors. The category depends on proximity (Taşçı & Onan, n.d.).

Artificial neural networks: Systems that can reveal new results from what has been learned thanks to the learning function are called. These networks, which are used in many fields, are established through the communication and relationship between neuron cells (Yıldırım, 2020).

Decision Support Machines: It is preferred to separate the connections between the data suitable for two different classes in the best possible way (Ülgen, 2017).

Time series analysis studies the persistence or behavior of information. It can be used for all kinds of timekeeping. It can be used to predict what will happen after learned events (Wikipedia, 2020).

After the Prediction Methods Models section, there are descriptive models: Clustering is the process of grouping related types and features that have not been classified in a cluster before. When evaluating the features of these juxtapositions, it turns out that they are most similar or share the same features (Ada et al., n.y.).

Analysis of relationships: examining relationships between processed products. This technique is typically used in shopping cart analytics. (Sabah & Bayraktar, 2020).

Sequence analysis is the study of events that occur one after the other. The application of this procedure is useful when the steps of an event are present and occur in sequence (Aytaç & Bilgin, t.d.).

Outlier analysis: The data determined in a data set is a completely different, crazy data. (ISTMER, n.y.).

Among these models, which are under two main headings as foresight methods and descriptive methods, suitable models are selected in line with the needs of the study to be carried out. For this reason, in the selection of the model, the suitability of the model for the needs is taken into account.

Literature Overview

Erkurt and Yıldırım collaborated with the R program to create infographics that facilitate housing selection, provide information about certain criteria of housing, and make comparisons between provinces/districts. They carried out the study using the data set obtained from the advertisements in the provinces of Istanbul, Ankara and Izmir. It has been concluded that a method has been developed that provides user satisfaction with the visual experience that people make between advertisements with the created infographics (Erkurt and Yıldırım, 2021). Thus, with this study, which increases user satisfaction, it has been seen that a visual source can be obtained to show the differences between the houses with the help of graphics.

In this study, Xiao et al. exploring potential applications of Data Mining techniques for effective use of large building operations data. They worked with decision tree and association rule methods from estimation techniques. The aim is to obtain information to identify building models. As a result of this study, useful results were obtained for the value of the building cooling load and the targets were achieved (Xiao et al., 2017). The study revealed the estimation values of the building cooling load with new building models.

Uzut and Buyrukoğlu wanted to predict real estate prices by applying and comparing data mining algorithms. Choosing linear regression, random forest, and gradient enhancement algorithms, they built models using these three algorithms and compared the performance of all. As a result, they concluded that the gradient enhancement algorithm provides the highest performance in real estate price estimation (Uzut and Buyrukoğlu, 2020). As a result of the study, the aim of finding a solution to the problem of correct pricing of real estate has been achieved.

Sandali Khare et al. performed a pricing calculation for buyers to estimate the price of the house. They stated that house price estimation will predict house prices according to various characteristics. When all algorithms and models that can be used in practice are compared, it is concluded that polynomial regression, a model that includes all 18 features, gives the best results. It was concluded that this would be more successful in predicting home prices for buyers (Khare et al., 2021). Calculation of house prices with many features and price estimation.

Baldauf et al. This study examines the impact of climate change on the long-term risks of home price valuations in the United States. As a result of the analysis, they found that the variability of belief in climate change risk is reflected in the evaluation of firms' estimates of real estate prices (Baldauf et al., 2020). The effect of climate on home valuation has been observed.

Dettling and Kearney examined how current house prices affect births in the current period. Here they worked with the coefficients obtained from the regression analysis. The results show that rising house prices are leading to a Slowing of births among non-homeowners and a Slowing of births among homeowners. However, they felt that this analysis was not designed to have definitive coverage (Dettling & Kearney, 2013). It has been concluded that house prices affect birth rates.

In this study, Huang and Mao conducted a study on marketing method for private companies using data mining. As a result of the study, they divided the customers into two subgroups according to their shopping areas as large and small. For this reason, it has been suggested that the employee can make predictions by monitoring customer data (Huang and Mao, 2022). This will make it easier for potential customers to choose.

Badriyah et al. aims to shorten the housing search process. A suggestion system was developed for the residence, where the feature requested by the user is provided. In addition, the search time should be minimized. People who do not have any special needs were selected as the target group. A suggestion system was created with content-based filtering method for apartment search. By doing this, the system shows that it is able to offer suggestions suitable for user preferences (Badriyah et al., 2018). It has been ensured that the system can present the desired house to the users in a short time.

Many contributions have been made with these studies. Especially with many studies in the field of housing price estimation, developments continue day by day. In addition, it is investigated which factors are effective in house pricing and the results emphasize various factors. Applications aimed at customer satisfaction and attracting customers are also very important.

The method to be used in this study is related to previous studies as it will perform estimation. In line with these positive results seen in estimation studies, the use of estimation method was encouraged. Previous studies have focused on various factors and analyzed whether there is a relationship. In this study, the relationship with the demographic characteristics of the tenants will be analyzed. In some studies, there is the aim of customer satisfaction, and in this study, it is aimed that the landlords are satisfied with their tenants. This means that they are also satisfied with their real estate agents.

Data Mining Success Factors and Challenges in The Real Estate Sector

The use of data mining in the real estate sector can bring many benefits as well as some difficulties. Here are the success factors and challenges when using data mining in real estate:

The factors that determine the success of data mining in the real estate sector are the quality and diversity of the data, the suitability of the selected model, and the predictive analysis ability. Data mining is based on accurate, complete and up-to-date data. Data quality is the key to success in real estate. It is important to apply data validation processes to access the right data sources and to minimize data entry errors in order to ensure that the data is of high quality (Koçak and Ergün, 2023).

Data mining in real estate requires a wide range of data, including different data types (data diversity). Leveraging different data sources such as real estate prices, location data, market trends and customer preferences leads to more effective results (IBM, n.d.). Correct model selection is important for the success of data mining projects. Common data mining techniques used in real estate include regression analysis, time series analysis, clustering and classification. Choosing the right model increases the accuracy of the predictions and the reliability of the analysis results (American Statistical Association, 2022). Predictive analytics skills are essential in data mining projects. Being able to predict future trends, price changes and customer preferences based on historical data is a valuable asset in the real estate field (Baykal, 2006).

In this study, it is aimed to obtain the data as completely as possible by directly applying to the real estate agency, which is a primary source. Since these data were collected at the time of the study, they are up-to-date. The appropriate model was determined by considering all the requirements.

The factors that determine the difficulties of data mining in the real estate industry are the quality, accessibility, management, protection and security of data, and the implementation and interpretation of data are among the challenges. The quality and accessibility of the data used in the real estate sector can be limited. Problems such as missing or incorrect data or incompatibilities between data sources can occur. To overcome these difficulties, data collection, cleaning, extraction and integration processes need to be carefully managed (Koçak and Ergün, 2023). Thus, it is possible to increase the data quality with preprocessing. A large amount of data can be generated in the real estate sector. Managing this large amount of data can be difficult. This data can be difficult to process, store and analyze. Techniques and infrastructures for big data management play an important role in data mining projects (Atan, 2016). Large-scale data can be narrowed so that more comfortable work can be achieved. Real estate data often contains sensitive information and can raise privacy and security concerns. Adequate data protection measures should be taken in data mining projects and data security should be kept at the highest level (Bardak, 2018). It is important that the results of data mining analyzes are properly interpreted and translated into business decisions. This process requires the implementation of the analysis results and the creation of value (Hoş, 2022).

DESIGN OF A RENT PAYMENT ESTIMATION SYSTEM

This section details the analysis of the collected data, model selection (decision tree), data preparation processes, and discussion of the results, starting with the data collection method for designing the rent payment estimation system.

Method of Data Collection

The source of the most relevant data for the study was determined as the real estate agent, where this data can be obtained in the most accurate way. Since the necessary data is related to the tenants in the real estate sector, a real estate agent was visited to collect the data face to face and the study was explained.

During the visit and interview to the Oked Real Estate office in the Karadeniz Ereğli District of Zonguldak, the aim of the research and application was explained to the consultant. Afterwards, tenant information was obtained with the help of the consultant. The reason for choosing this real estate agency was that it did a lot of work and contracted more tenants there. The data from here includes information on leases made in October, November, December 2021; January, February, March 2023; It was completed during 2022. Demographic information requested by the responsible consultant in the office during the study is gender, age, marital status, occupation, education level, number of people in the family, rent, number of children in the family, median income. and whether they pay the rent regularly.

The way to get the data was to review the contracts one by one, look at the information in the contract, and note down each tenant's information. While collecting the information, an Excel file was created and data was entered into this file as the information was collected. Then this Excel file is used in the application.

As a result of the studies carried out with the consultant, information was collected about 114 tenants in the above years. This created Excel file is then loaded into Rapidminer, the program used to run the data mining application.

The difficulty in the data collection phase was to manually enter Excel one by one, since the tenant data is kept by filing with a contract, not digitally.

Collecting the data one-on-one with the real estate agent in this way and obtaining the data directly in the written documents and contracts has increased the accuracy.

There were missing data and some unnecessary data in the obtained data. This was fixed by performing preprocessing before creating the model in the system. Missing data issue resolved. Some unnecessary fields have been removed.

Model Selection, Decision Tree Algorithm and Reason For Choosing the Algorithm

The dataset includes numeric, categorical, and binomial data. For this reason, it is necessary to have a model that can work with these three types of data. Additionally, it needed to be able to process incomplete data even if some of the data was missing. Not every model can work with missing data. It also had to be chosen from among prediction and classification models. Because as a result of the study, it was aimed to reveal an estimation.

For this reason, using the "Compare ROCs" function in the program, a comparison was made between the models meeting the above conditions and it was shown that the decision tree model was the best performing model among these models. For this reason, it was decided to use the Decision tree model (Rapidminer, 2017).

Demographic Data

In a real estate agency, information about tenants is usually kept by collecting the necessary documents together with the contracts and filing them for years. Access is not public as the files also contain some important personal information. Only real estate consultants can access this file. With the help of these files and the dialogue with the consultant in the form of questions and answers, the collected data was carefully processed and recorded in a file. The content of the created source of data is summarized in the following table.

9 demographic characteristics about tenants were obtained from the real estate consultant and a column was obtained on whether they should pay the rent regularly. This information contains two answers, yes or no.

Table 3.1 The Content of the Data Source

Gender (Binominal)	Age (Integer)	Marital Status (Polynominal)	Occupation (Polynominal)	Educational Status (Polynominal)
Male = 73	Under 30 = 39	Married = 66	Nurse = 8	License = 42
Female = 41	31-45 years old = 54	Single = 43	Government Employee = 11	Middle School = 15
	46 years and older = 21	Widow = 5	Worker = 32	Primary School = 15
			Pensioner = 7	High School = 28
			Engineer = 12	Associate Degree = 13
			Unemployed = 1	? = 1
			Housewife = 5	
			Tradesman = 26	
			Police = 3	
			Woman Teacher = 3	
			Manager = 2	
			Interpreter = 2	
			Male Teacher = 2	

Table 3.2 The Content of the Data Source 2

Number of people in the family (Numeric)	Rent (Integer)	Number of children in the family (Numeric)	Average income (Integer)	Do people pay the rent regularly? (Binominal)
1 = 39	Under 2000 = 27	0 = 81	Under 10000 = 29	Yes = 73
2 = 41	2000 – 4000 = 57	1 = 16	10000 – 20000 = 52	No = 31
3 = 19	4000 or more = 28	2 = 10	20000 or more = 31	
4 = 10		3 = 3		
5 = 2		4 = 2		
6 = 2		5 = 1		
7 = 1		? = 1		

In table 3.1: The data collected includes 73 male and 41 female tenants. The data includes 39 people under the age of 30, 54 people between the ages of 31 and 45, and 21 people aged 46 and over. The number of married people is 66, the number of single people is 43, and the number of widows is 5. Distribution by Occupation: the number of workers is 32, the number of unemployed is 1, the number of government employee is 11, the number of interpreters is 2, the number of tradesman is 26, the number of housewives is 5, the number of engineers is 12, the number of nurses is 8, the number of male teachers is 2, the number of woman teachers is 3, the number of pensioners is 7, the number of policemen is 3 and the number of manager is 2. The educational status and the number of people are as follows: 42 people have an associate degree. There are 15 people in primary school. 15 People are high school. 28 People have a license. There are 13 people in the middle school. This information of 1 person is not available in the table of data.

In table 3.2: In the family there are 81 who do not have children. the number of people living in a family with one child is 16. the number of families with two children is 10. A family with 3 children consists of 3. Family with 4 children is 2. A family with 5 children is only 1. The number of children missing in the table is 1. The number of tenants under the rental fee of 2000 is 27. Between 2000 and 4000 rents are 57 people. 4000 or more rentals belong to 28 people. The people who live alone are 39 people. The two living are 41 and 3 persons are 19. Living in the family 4 persons are 10 tenants. There are 2 tenants who live 5 people. 6 People live 2 tenants and 7 people live 1 tenant. If we consider the average income of tenants: the number of tenants receiving less than 10000 is 29. The number of tenants renting between 10000 and 20000 is 52. The number of tenants with an income of 20,000 or more is 31.

Gini Index

The parameter to be used in the decision tree model has been selected. The Gini Index measures the homogeneity of classes. The homogeneity is highest if all the data belong to a single class. The formula for the Gini Index is given here.

Gini = 1 - (p_1^2 + p_2^2+... +p_k^2)

The Gini index value is between zero and one. The property with the lowest Gini index is used to divide the data set. To start the division, the Gini index of all properties is calculated first. Then the property with the lowest value performs the division. After this separation, the classes can be differentiated again. In this case, the same actions are repeated to select the properties that perform the division. When the gini index is selected in the program, a tree is created by the system at the back (MK, 2020).

System implementation

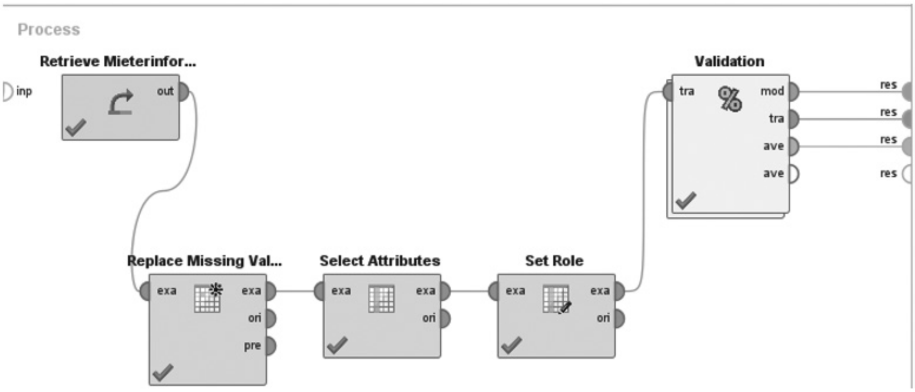


Figure 3.1 Model Process

Here, all the processes of the model created in the program are included.

Data preparation processes: Before performing a study on the data in the source, this data must be correct. Because it is necessary to reach a reliable conclusion. In addition to being complete, unnecessary data should not be included. First, the source of the data was loaded into the system from the computer using import data. Since some variables in the source of data are missing data, the process has been expanded to include the "Replace missing values" function used for this.

Then the subset was selected from the Attribute Filter Type section. The subset only helped select variables with missing data. In the Attributes part, the variables with missing data are selected from the Select Attributes part. Then the Select Attributes feature was added to the process. Inclusion attributes are selected in the Type section of settings. This indicates that the attributes to include are being selected. The subset is selected in the Attribute Filter Type section. Certain attributes have been selected. These are the properties that should be present in the decision tree and are believed to have an impact on the splits. Added the Set Role feature to the process. The reason for this is that the target value has to be determined. From edit list section " Do they pay the rent regularly? variable is set to label. Then the split validation feature was added. This feature separates the data in the source of data into training data and test data. A value of 0.7 was determined here after the tests. The section automatically specifies how to separate the data.

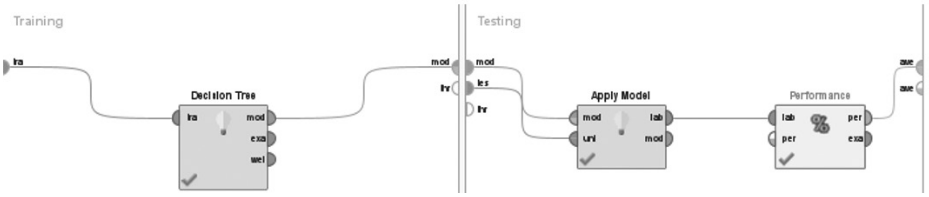
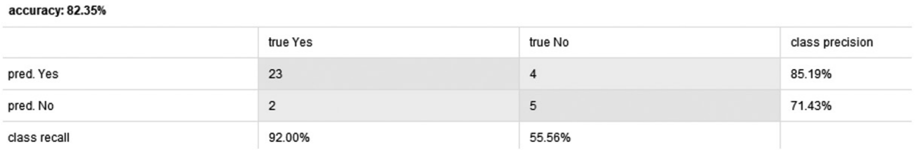


Figure 3.2 The process inside the validation operator

It is inside the "Validation" operator. Training and testing on the model is performed here. Here, the model trained with the Apply model function is tested with test data. Estimates were made. Finally, the calculation of the performance of the realized model was required and the "Performance" operator was added. The model is now complete.

Study Results



accuracy: 82.35%

	true Yes	true No	class precision
pred. Yes	23	4	85.19%
pred. No	2	5	71.43%
class recall	92.00%	55.56%	

Figure 3.3 Confusion Matrix

Accuracy 82.35% determined. This ratio indicates the overall accuracy of the estimate. In the generated table, the estimates made for the target variable are presented numerically. The variable contained two classes: yes and no. The data divided into these classes is displayed as the result of the estimation.

23 yes samples were correctly predicted, and 2 of them were predicted as no instead of being predicted as yes. 5 no samples were correctly predicted and 4 were predicted as yes instead of no.

Precision and recall values are calculated to ensure the accuracy of the study.

Precision shows the ratio of positively predicted values to correctly defined positive values.

Precision = (True Yes Pred Yes)/[(True Yes Pred Yes)+(True Yes Pred No)] = 23/(23+2) = 0.92

Sensitivity (recall) is a measure of how many transactions that need to be correctly estimated are correctly predicted.

Recall = (True Yes Pred Yes)/[(True Yes Pred Yes)+(True No Pred Yes)] = 23/(23+4) = 0.8519

Looking at these two metrics is not enough to make a decision about model performance. In addition, the F1 score should be calculated from these two measures.

The F1 score is a measure expressed as the harmonic mean of the precision and recall scores.

It was concluded that F1 = 2*[(precision*recall)/(precision+recall)] = 2*[(0.784)/(1.772)]=2*0.442=0.89 with a harmonic mean.

The reason for calculating the F1 score is to prevent erroneous results in data sets that are not evenly distributed. The F1 score can provide a more informative performance evaluation where the measure of accuracy is insufficient or may be misleading. The closer the F1 score is to 1, the better-matched accuracy and precision. In the study, it was found to be 0.89. Thus, it can be said that a proper relationship has been achieved.

The rules involved in building the decision tree are as follows:

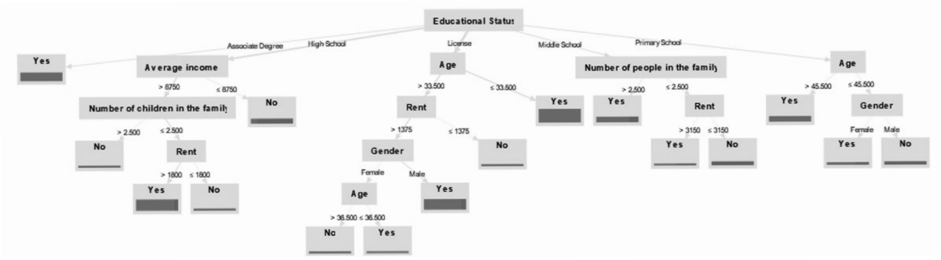


Figure 3.4 Decision Tree

When the decision tree formed in the system is examined, it has been found that the most important criterion is education status. Because it was chosen as the criterion for starting the tree in Gini index calculations. Then, because the target variable values of this variable were not exactly the same, a division was performed again. Here, a calculation is made among other variables. The gini index was calculated for each attribute used and the decision tree took its final form with division operations. Now the tree formation is complete when the target variable values are exactly the same for that variable.

Discussion of Study Results

An overview of the results shows that the most important criterion in the decision tree, based on what (santhoshini21k, 2021) defines, is educational status. Then, by dividing by the root node and performing calculations, it can be said that the transaction was performed with the property that is more important to the sample data. Demographics, occupation, number of children in the family characteristics were not used in the decision tree in this study. Therefore, it can be said that these features are not distinguishing features. As already mentioned in (Ulgen, 2017), it can be said that with more data and criteria, more separations are made. Which states that decision trees end when the samples in the division all belong to the same class and there is no qualification to divide the samples (mail.baskent.edu.tr, n.d.). In this direction it can be said that the decision tree performs divisions on samples belonging to different classes and there are no other properties to divide. It ends successfully by doing all the divisions and making the final division. As (Huddar, 2022) mentioned in study, it can be said that the premature termination of the rule in some outcomes in the decision tree is due to the similarity between the data and the lack of diversity, which should be shared since two nodes are enough in some outcomes. In line with the result obtained with the data obtained, the target variable of the associate degree graduates is the same as "yes". For this reason, tree division is terminated here without the need for a new branching. The data from the last node now returns the same result, and this is the only way the rule can be terminated. The data of the last node are exactly the same and do not differ from each other, now we can say that they give the same result and the rule can only be ended in this way. Thus, it can be said that the tree really ends with the completion of its operation.

According to previous literature research (Çalış et al., 2014) said that some of the questions asked in the survey were not a distinguishing feature in the decision tree and were not used. Demographics, occupation, and number of children in the family were not used in the decision tree in this study. It can be said that the reason for this is that these features are not distinguishing features and these features are not needed to get the result and therefore are not used in the rules. In subsequent studies, it can be said that if these features are replaced with other distinguishing features, a better tree can be built.

Considering such situations, comparison with the findings obtained in previous studies reveals similar situations.

CONCLUSION

The aim of the study was to predict whether tenants would pay the rent regularly. A sample of 144 tenants taken from the realtor was prepared. Here, estimation and classification methods and decision tree model are preferred. The study revealed that educated and relatively young people pay their rent regularly, while uneducated and young people do not pay their rent. In addition, it has been determined that the criteria of occupation and number of children do not have a direct relationship with rent. At the same time, it has been determined that the most important criterion is the level of education. Accordingly, real estate agents should pay attention to the education level, age or average income of the people when renting a house. Considering the characteristics of potential tenants, it will be possible to make estimations with the help of the application developed on those tenants. Real estate agents will be able to adapt this developed system easily with the existing tenant information in their database and will be able to get help from new tenants in choosing the most suitable one. It will also be able to learn directly by comparing the information of a tenant who is likely to make a contract, from the system. The aim was to emphasize that some demographic characteristics have a direct impact on people's regular rent payments. Research results also support this. The demographics discussed here are the ones that should be considered in potential tenants. With this study, it is possible to achieve an improvement in rent payments in the real estate sector. When real estate agents implement this study, they will gain a more reliable image among homeowners and their volume in the market will increase. Negotiating with a tenant who is expected to pay the rent properly is economically advantageous for both the real estate agent and the landlord. The regular income of the landlords will continue. Thanks to regular rent payments, mutual relations will also be able to progress in a healthy way. Trust will increase between the landlord and the real estate agent and the relationship will progress in a positive direction. The study was carried out only with the data of 114 tenants in a single agency (OKED Emlakçı) in KDZ.Ereğli district. Diversity is also low, as the data covers tenants in a particular region. For this reason, the area where the data is collected can be expanded, which is recommended for future studies, or the diversity can be increased by bringing together data from different institutions. In other words, it is recommended to work with much more data than 114 data and to carry out a wider data collection process for future studies. Thus, the diversity of data will increase and will positively affect the percentage of predictions. In addition, it is possible to seek ways to develop a model with better performance. Or, if a different source is used as a data source, it may be possible to obtain better results. The data in this study were obtained with the help of OKED Real Estate Agents in Zonguldak-Ereğli. Gratitude is required.

REFERENCES

Ada, M., Altunay, F., Civelek, M., Kaplan, S., & Koç, P. (n.d.). *Kümeleme Analizi*. tip.baskent.edu.tr. Retrieved April 13, 2023, from http://tip.baskent.edu.tr/kw/upload/464/dosyalar/cg/sempozyum/ogrsmpzsnm13/13.P9.pdf

Akay, M., Kaya, Z. S., & Günkut, M. (2023, March 22). *Veri Madenciliği: Konut Fiyat Endeksi ve Konut Satış Sayılarının Basit Doğrusal Regresyon Analizi*. DergiPark. 1309-8020, 25-44.Retrieved April 14, 2023, from https://dergipark.org.tr/tr/download/article-file/2915783

American Statistical Association, (2022, February 1). *Ethical guidelines for statistical practice*. American Statistical Association. https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice

Atan, S. (2016, May 25). *Veri, Büyük Veri ve İşletmecilik*. Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi. 35(19), 137-153. https://dergipark.org.tr/tr/download/article-file/852532

Aytaç, E., & Bilgin, T. (n.d.). *Sıralı Örüntü Madenciliği Yöntemi Kullanılarak İnternet Bankacılığı Kullanıcı Davranışlarının Modellenmesi*. ab.org.tr. Retrieved April 13, 2023, from https://ab.org.tr/ab14/bildiri/134.pdf

Badriyah, T., Azvy, S., Yuwono, W., & Syarif, I. (2018). *Recommendation System for Property Search Using Content Based Filtering Method*. 2018 International Conference on Information and Communications Technology. 60111. 25-29. ieeexplore.ieee.org. https://ieeexplore.ieee.org/author/37086000175

Baldauf, M., Garlappi, L., & Yannelis, C. (2020). Does climate change affect real estate prices? only if you believe in it. *SSRN Electronic Journal*, 1256–1295, 1-53. https: //deliverypdf.ssrn.com/delivery.php?ID = 1851190050710110061061040721180270740170 47006041059002118104074106085111026025096112025016 100118110061032000117007112108094077016080011050064108083026066083125088094025044095 0680961 200700840921161190721221070060651190870060860291160261061130060 17093098&EXT =pdf&INDEX=TRUE

Bardak, T., & Sözen, E. (2018, October). *Veri Madenciliğinin Önemi*. Journal of Business Economics and Management Research (2). 21-29. ttps://personel.omu.edu.tr/docs/ders_dokumanlari/9273_24732_1390.pdf

Başkent Üniversitesi (n.d.). *Karar Ağacı (decision Karar Ağacı (Decisiontree)))) Nedir?*. mail.baskent.edu.tr. https://mail.baskent.edu.tr/~20410964/DM_8.pdf

Baykal, A. (2006). *Veri Madenciliği Uygulama Alanları*. dergipark.org.tr. 7, 95-107. https://dergipark.org.tr/tr/download/article-file/787239

Çalış, A., Kayapınar, S., & Çetinyokuş, T. (2014, September 10). *VERİ MADENCİLİĞİNDE KARAR AĞACI ALGORİTMALARI İLE BİLGİSAYAR VE İNTERNET GÜVENLİĞİ ÜZERİNE BİR UYGULAMA*. “ Endüstri Mühendisliði Dergisi. 25(3,4), 2-19. https://dergipark.org.tr/tr/download/article-file/752270

Çelik, M. (2009). *Veri Madenciliğinde Kullanılan Sınıflandırma Yöntemleri ve Bir Uygulama*. nek.istanbul.edu.tr. Retrieved April 14, 2023, from http://nek.istanbul.edu.tr:4444/ekos/GAZETE/index.php

Dettling, L., & Kearney, M. S. (2013). House prices and birth rates: The impact of the real estate market on the decision to have a baby. *Journal of Public Economics*. 1050,1-54. https://doi.org/10.3386/w17485

Erden, C. (2020, August 12). *9- Regresyon Analizi ve Bir Uygulama*. YouTube. Retrieved April 14, 2023, from https://www.youtube.com/watch?v=vNeGFl140m4

Erkurt, E., & Yildirim, E. (2021). Bir Büyük Veri Görselleştirme Uygulaması Olarak Konut Tercih İnfografikleri. *İktisadi ve Idari Bilimler Fakültesi Dergisi*. 23(1), 37-52. https://dergipark.org.tr/tr/download/article-file/1090959

Hoş, S. (2022, September 2). *Veri Madenciliği (data mining) Nedir?* ÇözümPark. https://www.cozumpark.com/veri-madenciligi-data-mining-nedir/

Huang, R., & Mao, S. (2022). Research on precision marketing of real estate market based on Data Mining. *Scientific Programming*, 8198568, 1–13. https://doi.org/10.1155/2022/8198568

Huddar, M. (2022, January 29). *Build decision tree using Gini Index solved numerical example machine learning by dr. Mahesh Huddar*. YouTube. https://www.youtube.com/watch?v=zNYdkpAcP-g

Huria, R. (2014, October 22). *DM - the most important thing you probably aren’t using*. Loginworks. Retrieved April 7, 2023, from https://www.loginworks.com/blogs/217-data-mining-and-its-importance/

IBM, (n.d.). *Büyük Veri Analitiği*. IBM. https://www.ibm.com/de-de/analytics/hadoop/big-data-analytics

İSTMER, (n.d.). *AykIRI Değer Tespitinde Aritmetik Ortalama Ve Medyan Değerlerinin İncelenmesi*. İSTMER. Retrieved April 14, 2023, from https://www.istmer.com/aykiri-deger-tespiti-ortalama-medyan/

Khare, S., Gourisaria1, M. K., Harshvardhan1, G., Joardar2, S., & Singh3, V. (2021). *IOPscience*. IOP Science. Open Acces proceedings Journal of Physics. 1099, 1-14. https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012053

Koçak, A. & Ergün, M. A. (2023). *Sağlıkta Veri Kalitesi ve veri Madenciliği Uygulamaları*. Sağlıkta veri kalitesi ve veri madenciliği uygulamaları. dergipark.org.tr. 3(1), 23-30. https://dergipark.org.tr/en/download/article-file/2596222

Kodedu, (2014, May 28). *Naïve Bayes Sınıflandırma Algoritması*. KodEdu. Retrieved April 14, 2023, from https://kodedu.com/2014/05/naive-bayes-siniflandirma-algoritmasi/

MK, G. (2020, October 28). *Gini index for decision trees: Mechanism, perfect & imperfect split with examples*. upGrad blog. https://www.upgrad.com/blog/gini-index-for-decision-trees/

Rapidminer, Inc. (2017, September 21). *Finding the right model | rapidminer*. YouTube. Retrieved May 7, 2023, from https://www.youtube.com/watch?v=C8Ko3-2f-pA&list=PLssWC2d9JhOZLbQNZ80uOxLypglgWqbJA&index=16

santhoshini21k, (2021, January 19). *Understanding the gini index in decision tree with an example*. Numpy Ninja. https://www.numpyninja.com/post/understanding-the-gini-index-in-decision-tree-with-an-example

Taşçı, E., & Onan, A. (n.d.). *K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi*. ab.org.tr. Retrieved April 13, 2023, from https://ab.org.tr/ab16/bildiri/102.pdf

Ulgen, K. (2017, November 12). *Makine Öğrenimi Bölüm-5 (Karar Ağaçları)*. Medium. Retrieved April 30, 2023, from https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimi-b%C3%B6l%C3%BCm-5-karar-a%C4%9Fa%C3%A7lar%C4%B1-c90bd7593010

Ulgen, K. (2017, October 30). *Makine Öğrenimi Bölüm-4 (Destek Vektör Makineleri)*. Medium. Retrieved April 14, 2023, from https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimi-b%C3%B6l%C3%BCm-4-destek-vekt%C3%B6r-makineleri-2f8010824054

Uzut, Ö. G., & Buyrukoğlu, S. (2020). *Veri Madenciliği Algoritmalari ile Gayrimenkul Fiyatlarının Tahmini*. euroasiajournal.org. https://euroasiajournal.org/index.php/ejas/article/view/41/39

Wikipedia, (2020, December 6). *Zaman Serisi*. Wikipedia. Retrieved April 14, 2023, from https://tr.wikipedia.org/wiki/Zaman_serisi

Xiao, F., Wang, S., & Fan, C. (2017, May). *Mining Big Building Operational Data for Building Cooling Load Prediction and Energy Efficiency Improvement*. 2017 IEEE International Conference on Smart Computing. 978-1-5090-6517-2 ieeexplore.ieee.org. https://ieeexplore.ieee.org/document/7947023

Yıldırım, E. (2020, May 2). *Yapay Sinir Ağı (artificial neural network) Nedir?* Veri Bilimi Okulu. Retrieved April 14, 2023, from https://www.veribilimiokulu.com/yapay-sinir-agiartificial-neural-network-nedir/#:~:text=Yapay%20sinir%20a%C4%9Flar%C4%B1%20(YSA)%2C,geli%C5%9Ftirilen%20bilgisayar%20sistemleridir%5B1%5D.

Zülfikar, H. (2022). *Veri / data*. Sosyal Bilimler Ansiklopedisi, 978 - 605 - 312 - 432 – 0. Retrieved April 7, 2023, from https://ansiklopedi.tubitak.gov.tr/ansiklopedi/veri_data